External documentation

This file includes a documentation of the internal data structures used in our project.

1. **Main – Project Main Script**
   A script that defines a command line parser and runs all the different experiments.
   (To run experiments in the project please read Experiment_Guide.txt)

2. **Data – Project Data**
   A data folder that contains 40 files of FCS data, a file for each patient.
   Each file contains thousands of single-cell datasets measured by CyTOF.

3. **Classifications – Project Data**
   A data folder that contains an annotation file which includes the patients TIL/ACT classifications (responder/ non-responder)

4. **Citrus_data – Project Data**
   A data folder that contains Citrus data file that includes the abundances of patients' cells in Citrus clusters.

5. **Main functions – Project Auxiliary Functions**
   A script that defines commonly used functions for patient's data manipulations and helper functions to perform the clustering algorithm.
   The script includes functions related to sampling cells, translating data files into readable formats, running cross validation over classifiers, processing clusters data and so on…

6. **Entities – Project Objects**
   A package that defines 3 main entities:
   1. **Human being**
      a class that represents a patient and it stores patient's cells with its file index.
   2. **Node**
      a class that represents a single node in the clustering tree, it stores a populations of cells that was obtained from the parent node, and details regarding its position in the tree.
   3. **Tree**
      A class that represents the clustering tree, it stores the minimal entropy/cells parameters, the root node as well as the paths constructed recursively during the clustering procedure.

7. **Samplers – Sampling Procedures**
   A package that contains two different methods for sampling patients' cells randomly.
   1. **Sample By Size**
      Sampling fixed number of cells from each patient.
   2. **Sample By Percentage**
      Sampling amount of cells from each patient by percentage.

8. **Calc_thresh – Markers Thresholds Functions**
   A package that contains auxiliary functions to calculate markers distance thresholds and entropy values.
   1. **R_scripts**
      A script written in R which contains a non-deterministic mixtool function which finds the distance thresholds given a dataset of cells and markers values, using a normal mix model which depends on the Gaussian distribution and k-mean threshold.
   2. **Calc_thresholds**
      A python wrapper for the function from **R_scripts**.

9. **Helpers – Organizing Markers Functions**
   A package that contains methods for storing intermediate data regarding markers indices and names using pickle objects.

10. **Classifiers – Project Classifiers**
    A package that contains a decision-tree classifier we built to analyze the benefits of the received clusters from the constructed algorithm. The classifier can be used to predict TIL/ACT success over a melanoma patient.

11. **Builders – Classifier Builders**
    A package for classes that encapsulates the creation of a classifier.
    1. **Cart_builder**
       A class that encapsulates the creation of a CART (classification and regression tree) classifier.

12. **Validators – Classifier Validators**
    A package that contains different methods for testing the effectiveness of predictions, using cluster's information.
    1. **Regular cross-validation.**
    2. **Randomly split cross-validation**.

13. **Experiments – Project Experiments**
    1. **New_cluster_experiments - The Clustering Algorithm Experiments**
       A script which contains the main clustering algorithm constructed in the project. The algorithm begins by randomly sampling a number of cells from each patient, then saving the all the sample data in the **sample_dir** folder.
       The algorithm then recursively builds the clustering tree given the minimal cells parameter and the minimal entropy parameter. When the procedure ends, all the clustering output files are saved under the **results** folder, including the patient-cell abundances matrix, which is later used with the corresponding patient labels to train the cart classifier. Finally we estimate the classifier performance by using the cross validation technique and then saving the rates as well, under the **results** folder.
    2. **Citrus_experiments – Citrus Experiment**
       A script that estimates the classifier's performance when trained over data of Citrus.

    To run one of these experiments in the project please read section 1 and section 2 in Experiment_Guide.txt. There you can find an explanation on how to conduct the two experiments in detail.

14. **Process_results – Collecting Results**
    A script that summarizes all the experiments results nicely.
    If you have conducted clustering experiments in the project: please read section 1.B in Experiment_Guide.txt to understand how to summarize your experiments results.

15. **Analyze_results – Clusters Visualizations**
    A script that analyzes the results by performing T-SNE visualizations of the results.
    If you are interested in observing our clusters 2D visualizations: please read section 3 in Experiment_Guide.txt to understand how to perform a visualization experiment.

16. **Graph_best_results – Project Graphs**
    A script that contains functions for plotting 2D linear graphs demonstrating the experiments' results.

17. **Sample_dir – Samples Directory**
    A folder that contains pickle files which store the sample-cells when a sampling process is performed during the clustering algorithm.
    The folder is empty when no clustering experiments are conducted.

18. **Results – Project Results**
    1. **Citrus_results**
       A folder that stores files related to Citrus, including the abundances matrix and the classifier prediction rates.
    2. **Clustering_algorithm_results**
       A folder that contains experiments folders which includes all the output files saved after performing the clustering (E.g. tree files, classifier results abundances matrix etc…).
       The folder is empty when no experiments are conducted.
    3. **Experiments_summary**
       A folder that contains a summary tables of all the classifier's results of all experiments conducted under **Clustering_algorithm_results** folder.
    4. **Graph_results**
       The folder contains graphs of our experiments' results.
    5. **Tsne_results**
       The folder contains experiments folders which include output files saved after performing the analyze process such as visualizations graphs saved in PDF files.
       The folder is empty when no experiments are conducted.